DML-CZ – zpracování článků z retro-born-digital období[1]

<http://www.dml.cz/>

Michal Růžička

(1. ročník FI MU N-IN BIT, UČO: 143424, e-mail: <mruzicka@mail.muni.cz>)

16. prosince 2008

[1]Projekt byl podpořen grantem č. 1ET200190513 Akademie věd České republiky.

Projekt DML-CZ

- Česká digitální matematická knihovna.
- Vyvíjená od roku 2005.
- Cílem je uchovávání digitální podoby většiny matematické literatury, která byla kdy publikována na území českých zemí.
- Poskytován volný přístup.

Projekt DML-CZ

- Česká digitální matematická knihovna.
- Vyvíjená od roku 2005.
- Cílem je uchovávání digitální podoby většiny matematické literatury, která byla kdy publikována na území českých zemí.
- Poskytován volný přístup.

Projekt DML-CZ

- Česká digitální matematická knihovna.
- Vyvíjená od roku 2005.
- Cílem je uchovávání digitální podoby většiny matematické literatury, která byla kdy publikována na území českých zemí.
- Poskytován volný přístup.

Projekt DML-CZ

- Česká digitální matematická knihovna.
- Vyvíjená od roku 2005.
- Cílem je uchovávání digitální podoby většiny matematické literatury, která byla kdy publikována na území českých zemí.
- Poskytován volný přístup.

Tři hlavní období zpracovávaných časopisů

Tři hlavní období, se kterými se musí projekt digitální knihovny vypořádat:

- retro-digitalizační období – Dokumenty jsou dostupné pouze v papírové podobě a pro potřeby digitální knihovny musí být digitalizovány.
- retro-born-digital období – Dokumenty jsou již dostupné v elektronické podobě, ale byly připraveny bez ohledu na potřeby digitální knihovny. Formát těchto dokumentů je tak často nevhodný pro přímé vložení do digitální knihovny.
- born-digital období – Dokumenty jsou pořizovány elektronickou cestou takovým způsobem, aby byly uspokojeny jak požadavky vydavatele, tak potřeby digitální knihovny.

Tři hlavní období zpracovávaných časopisů

Tři hlavní období, se kterými se musí projekt digitální knihovny vypořádat:

- retro-digitalizační období – Dokumenty jsou dostupné pouze v papírové podobě a pro potřeby digitální knihovny musí být digitalizovány.
- retro-born-digital období – Dokumenty jsou již dostupné v elektronické podobě, ale byly připraveny bez ohledu na potřeby digitální knihovny. Formát těchto dokumentů je tak často nevhodný pro přímé vložení do digitální knihovny.
- born-digital období – Dokumenty jsou pořizovány elektronickou cestou takovým způsobem, aby byly uspokojeny jak požadavky vydavatele, tak potřeby digitální knihovny.

Tři hlavní období zpracovávaných časopisů

Tři hlavní období, se kterými se musí projekt digitální knihovny vypořádat:

- retro-digitalizační období – Dokumenty jsou dostupné pouze v papírové podobě a pro potřeby digitální knihovny musí být digitalizovány.
- retro-born-digital období – Dokumenty jsou již dostupné v elektronické podobě, ale byly připraveny bez ohledu na potřeby digitální knihovny. Formát těchto dokumentů je tak často nevhodný pro přímé vložení do digitální knihovny.
- born-digital období – Dokumenty jsou pořizovány elektronickou cestou takovým způsobem, aby byly uspokojeny jak požadavky vydavatele, tak potřeby digitální knihovny.

Tři hlavní období zpracovávaných časopisů

Tři hlavní období, se kterými se musí projekt digitální knihovny vypořádat:

- retro-digitalizační období – Dokumenty jsou dostupné pouze v papírové podobě a pro potřeby digitální knihovny musí být digitalizovány.
- retro-born-digital období – Dokumenty jsou již dostupné v elektronické podobě, ale byly připraveny bez ohledu na potřeby digitální knihovny. Formát těchto dokumentů je tak často nevhodný pro přímé vložení do digitální knihovny.
- born-digital období – Dokumenty jsou pořizovány elektronickou cestou takovým způsobem, aby byly uspokojeny jak požadavky vydavatele, tak potřeby digitální knihovny.

Schéma zpracování retro-born-digital časopisů

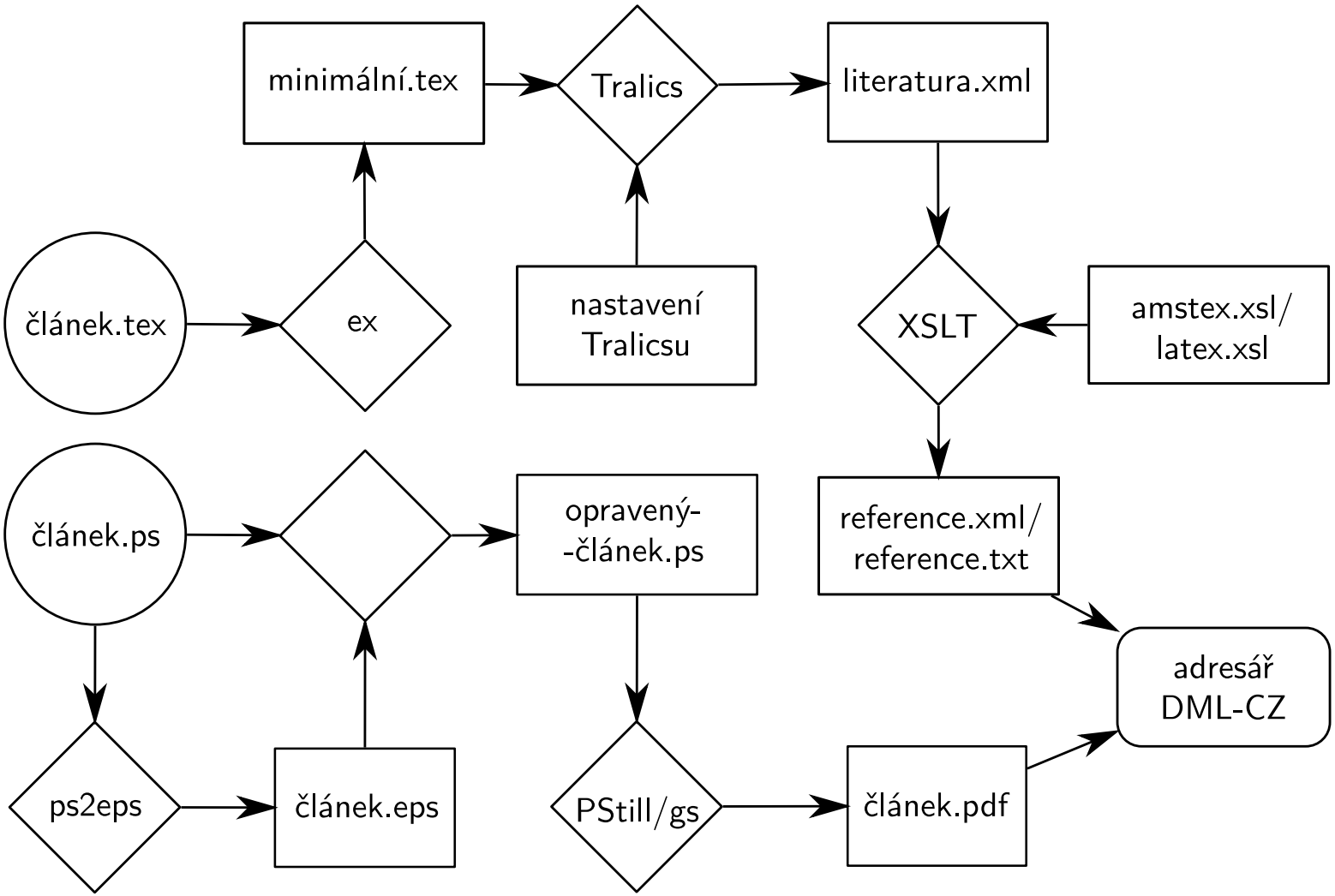


Schéma zpracování retro-born-digital časopisů

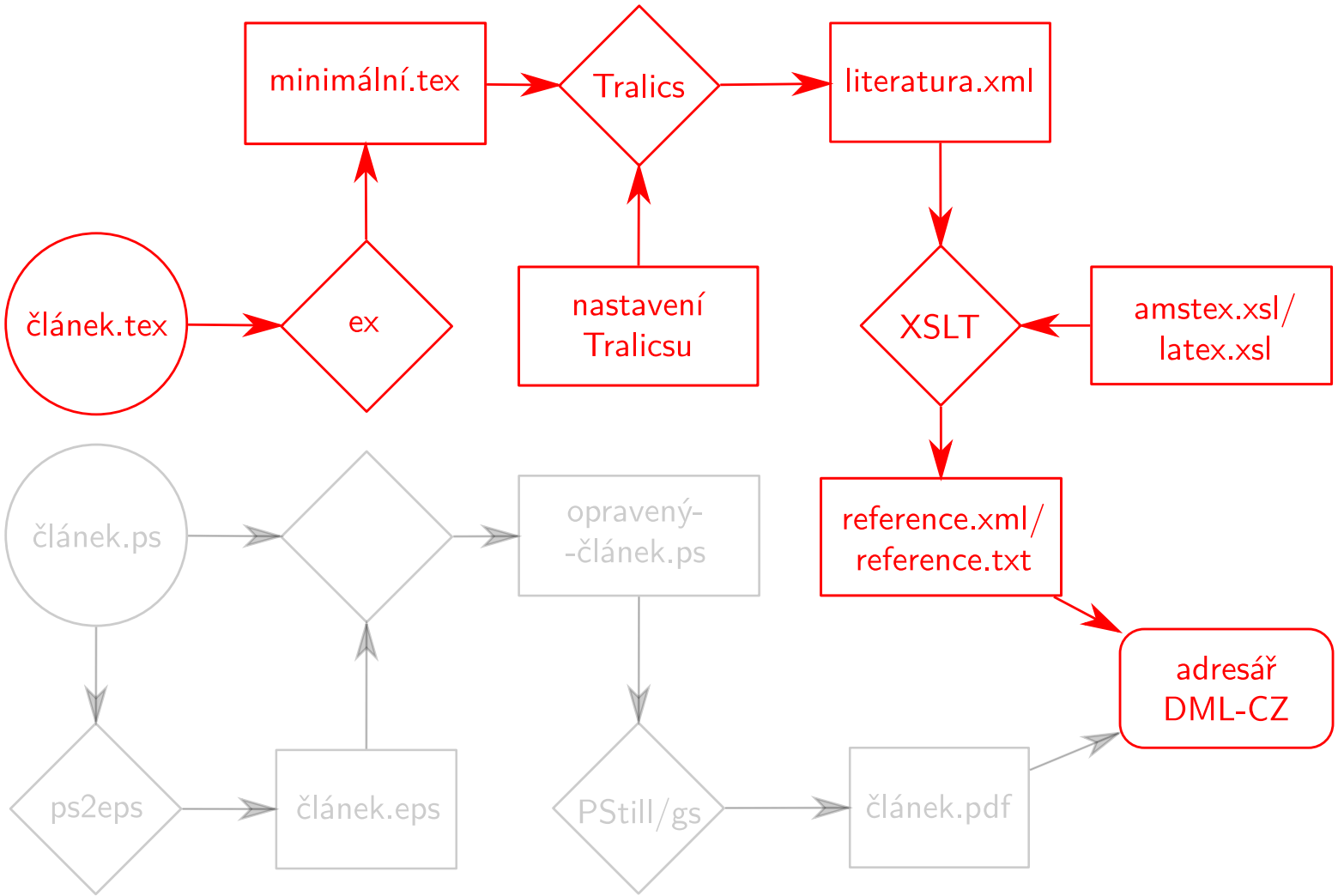


Schéma zpracování retro-born-digital časopisů

```
\documentclass{archivum}
\begin{document}
  \Refs
  \ref\key1
      \by Gancarzewicz, J., Michor P. W.
      \paper Natural...
  \endref
  \ref\key2
      \by Zajtz, A.
      \paper On the order of natural...
  \endref
  ...
  \endRefs
\end{document}
```

Schéma zpracování retro-born-digital časopisů

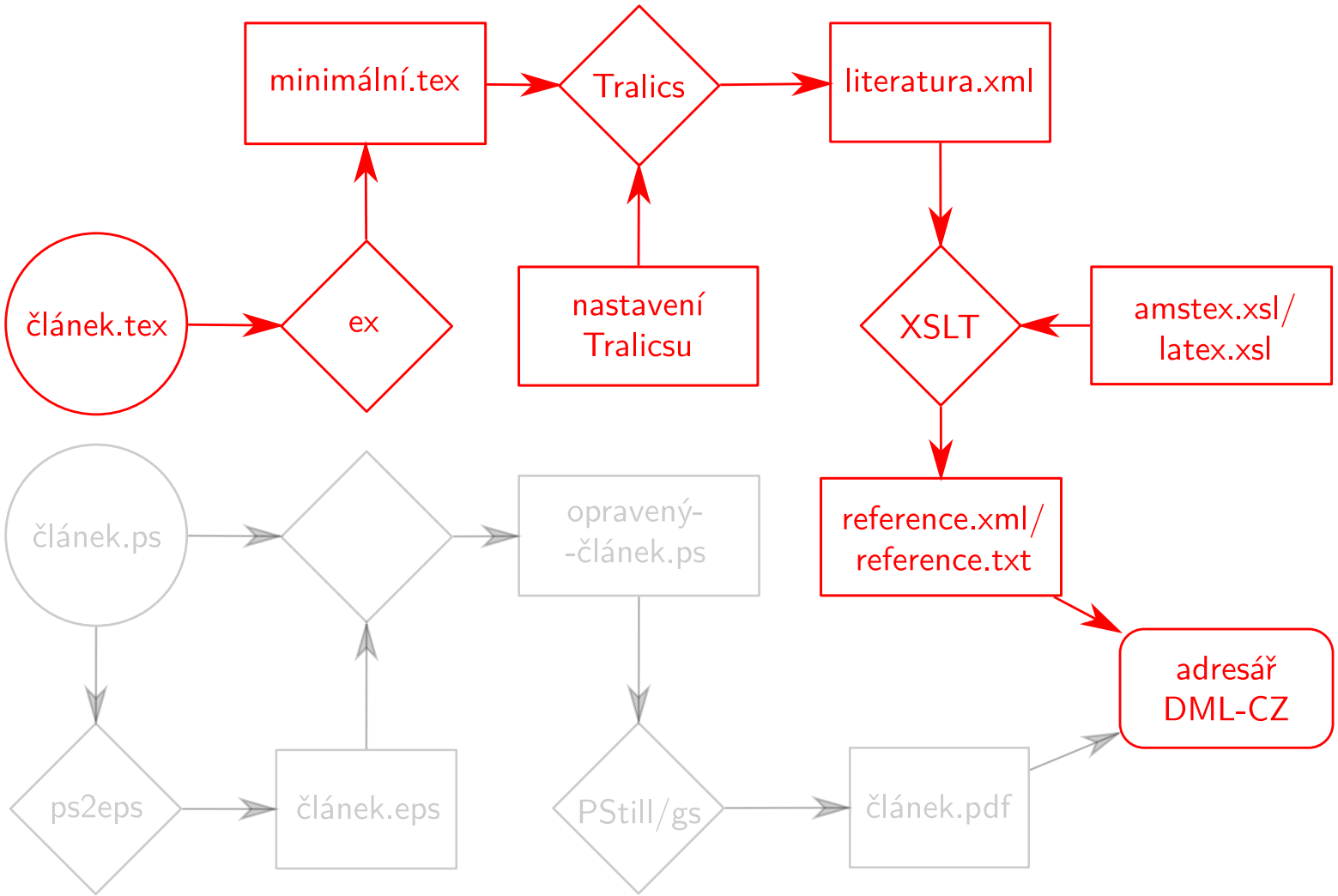
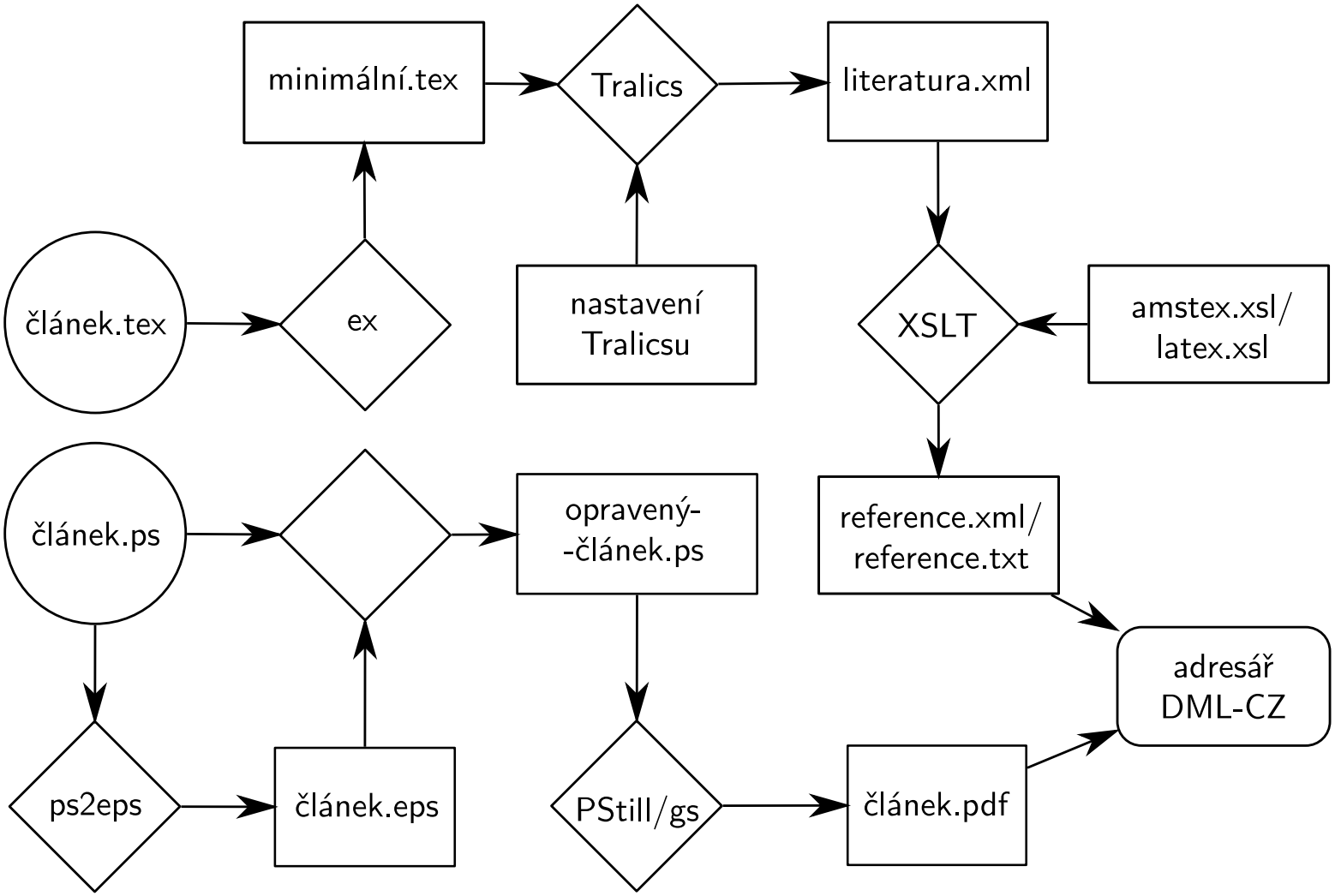


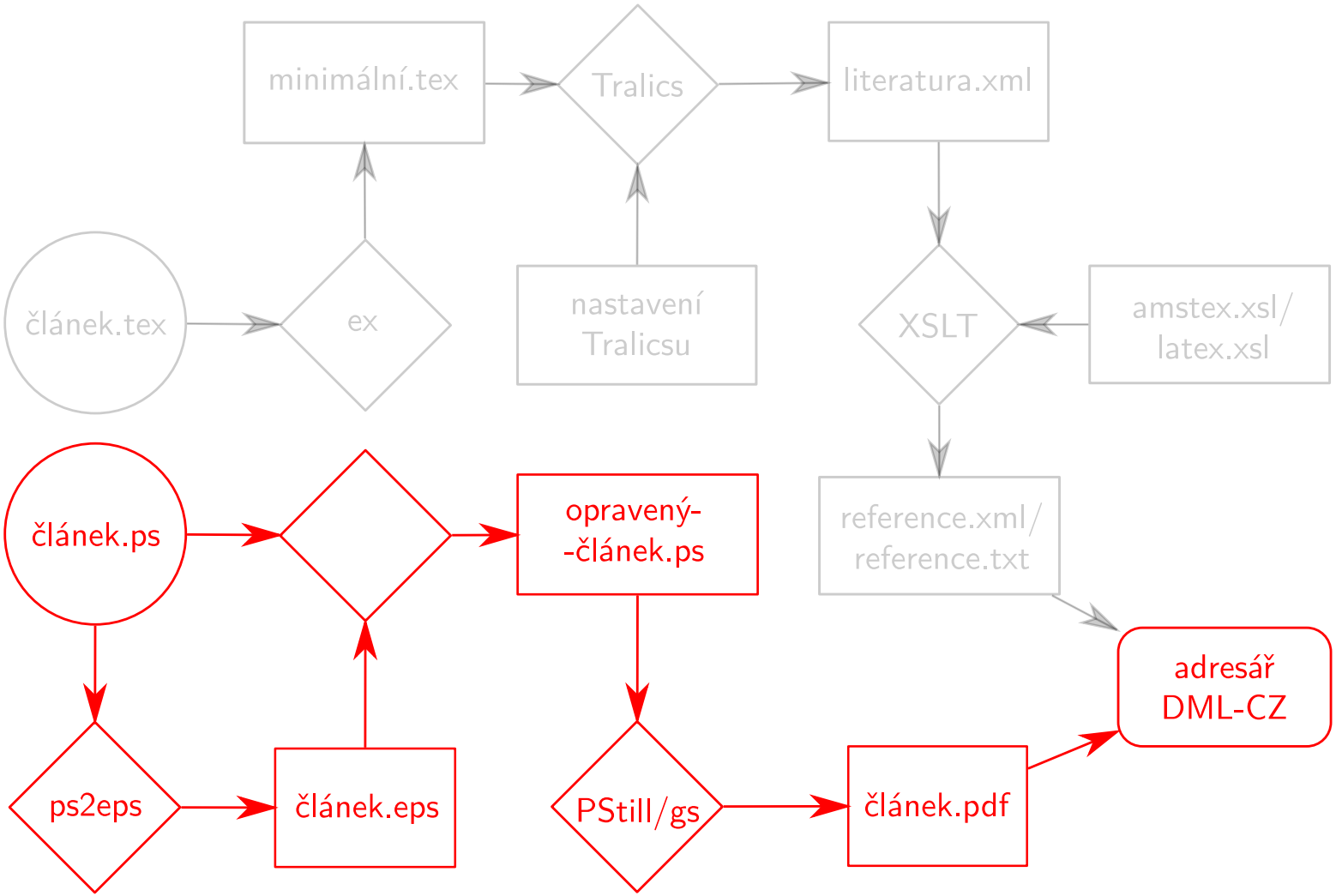
Schéma zpracování retro-born-digital časopisů

```
<?xml version="1.0" encoding="UTF-8"?>
<references>
  <reference id="1">
    <prefix>[1]</prefix>
    <title>Natural...</title>
    <authors>Gancarzewicz, J., Michor P. W.</authors>
    ...
  </reference>
  <reference id="2">
    <prefix>[2]</prefix>
    <title>On the order of natural...</title>
    <authors>Zajtz, A.</authors>
    ...
  </reference>
...
</references>
```

Náhrada bitmapových fontů



Náhrada bitmapových fontů



Náhrada bitmapových fontů

- Původní PostScriptové soubory obsahovaly bitmapové fonty v rozlišení jen 300 DPI.
- Pokus o náhradu původních fontů za fonty vektorové programem FixFont selhal.
- Částečný úspěch s programem PStill a náhradou fontů během konverze PostScriptů do PDF.
  - Závisí na metadatech vkládaných do PostScriptu během jeho vytváření programem `dvips`.

Náhrada bitmapových fontů

- Původní PostScriptové soubory obsahovaly bitmapové fonty v rozlišení jen 300 DPI.
- Pokus o náhradu původních fontů za fonty vektorové programem FixFont selhal.
- Částečný úspěch s programem PStill a náhradou fontů během konverze PostScriptů do PDF.
  - Závisí na metadatech vkládaných do PostScriptu během jeho vytváření programem `dvips`.

Náhrada bitmapových fontů

- Původní PostScriptové soubory obsahovaly bitmapové fonty v rozlišení jen 300 DPI.
- Pokus o náhradu původních fontů za fonty vektorové programem FixFont selhal.
- Částečný úspěch s programem PStill a náhradou fontů během konverze PostScriptů do PDF.
  - Závisí na metadatech vkládaných do PostScriptu během jeho vytváření programem `dvips`.

Náhrada bitmapových fontů

- Původní PostScriptové soubory obsahovaly bitmapové fonty v rozlišení jen 300 DPI.
- Pokus o náhradu původních fontů za fonty vektorové programem FixFont selhal.
- Částečný úspěch s programem PStill a náhradou fontů během konverze PostScriptů do PDF.
  - Závisí na metadatech vkládaných do PostScriptu během jeho vytváření programem `dvips`.

Shrnutí

- Získávání metadat přímo z původních (kvalitně označkovaných) zdrojových textů je přesnější a jednodušší než OCR.
- Zpracována retro-born-digital čísla časopisů Archivum Mathematicum, Acta Universitatis Palackianae Olomucensis a Applications of Mathematics.
- Data časopisů Czechoslovak Mathematical Journal a Mathematica Bohemica budou zpracována do konce roku.

Shrnutí

- Získávání metadat přímo z původních (kvalitně označkovaných) zdrojových textů je přesnější a jednodušší než OCR.
- Zpracována retro-born-digital čísla časopisů Archivum Mathematicum, Acta Universitatis Palackianae Olomucensis a Applications of Mathematics.
- Data časopisů Czechoslovak Mathematical Journal a Mathematica Bohemica budou zpracována do konce roku.

Shrnutí

- Získávání metadat přímo z původních (kvalitně označkovaných) zdrojových textů je přesnější a jednodušší než OCR.
- Zpracována retro-born-digital čísla časopisů Archivum Mathematicum, Acta Universitatis Palackianae Olomucensis a Applications of Mathematics.
- Data časopisů Czechoslovak Mathematical Journal a Mathematica Bohemica budou zpracována do konce roku.


Sojka, P.:
From Scanned Image to Knowledge Sharing.
In Tochtermann, K., Maurer, H., eds.: Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management, Graz, Austria, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co. (June 2005) 664–672.
ISSN: 0948-6968.


Bartošek, M., Lhoták, M., Rákosník, J., Sojka, P., Šárfy, M.:
DML-CZ: The Objectives and the First Steps.
In Borwein, J., Rocha, E.M., Rodrigues, J.F., eds.: CMDE 2006: Communicating Mathematics in the Digital Era.
A. K. Peters, MA, USA (2008) 69–79.
ISBN: 978-3-540-85109-7.


Sojka, P., Panák, R., Mudrák, T.:
Optical Character Recognition of Mathematical Texts in the DML-CZ Project.
Technical report, Masaryk University, Brno (September 2006) presented at CMDE 2006 conference in Aveiro, Portugal.


Bartošek, M., Krejčíř, V.:
Jak se dělá digitální matematická knihovna.
In Sborník konference AKP 2007, Liberec, Czech Republic (2007).
ISBN: 978-80-01-03691-4.
Available from WWW: <http://dml.muni.cz/docs/akp2007-sbornik.pdf>.


Czech Digital Mathematics Library [online].
[cit. 2008-05-30].
Available from WWW: <http://dml.cz/>.


Czech Digital Mathematics Library: About DML-CZ [online].
[cit. 2008-06-22].
Available from WWW: <http://dml.cz/about/>.

Shrnutí, reference

Archivum Mathematicum [online].
Masaryk University, Brno.
Last modified 14 May 2008 [cit. 2008-05-18].
Available from WWW: <http://www.emis.de/journals/AM/>.

Grimm, J.:
Tralics, a LaTeX to XML Translator.
In Proceedings of EuroTEX, TUGboat **24**(3) (2003) 377–388.
ISSN: 0896-3207.

Tralics: a LaTeX to XML translator [online].
Last modified $Date: 2008/05/13 09:32:16 $ [cit. 2008-05-18].
Available from WWW: <http://www-sop.inria.fr/apics/tralics/>.

TeX Live [online].
$Date: 2008/05/17 00:21:31 $ [cit. 2008-05-25].
Available from WWW: <http://www.tug.org/texlive/>.

Probets, S., Brailsford, D.:
Substituting outline fonts for bitmap fonts in archived PDF files.
Software-Practice and Experience. **33**(9) (2003) 885–899.
ISSN: 0038-0644.

Research - Fonts [online].
[cit. 2008-05-25].
Available from WWW: <http://www.eprg.org/research/fonts/>.

Siegert, F.:
PStill: ...generate, reprocess, normalize and extract content for PDF, EPS and PS. [online].
[cit. 2008-05-25].
Available from WWW: <http://www.pstill.com/>.

Shrnutí, reference

Applications of Mathematics [online].
Institute of Mathematics, Academy of Sciences of the Czech Republic.
Last changed January 23, 2007 [cit. 2008-12-05].
Available from WWW: <http://am.math.cas.cz/>.

Czechoslovak Mathematical Journal [online].
Institute of Mathematics, Academy of Sciences of the Czech Republic.
Last changed February 29, 2008 [cit. 2008-12-05].
Available from WWW: <http://cmj.math.cas.cz/>.

Mathematica Bohemica [online].
Institute of Mathematics, Academy of Sciences of the Czech Republic.
Last changed March 18, 2008 [cit. 2008-12-05].
Available from WWW: <http://mb.math.cas.cz/>.